

# KARMA: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking <sup>a)</sup>

Daryush D. Mehta,<sup>b)</sup> Daniel Rudoy, and Patrick J. Wolfe<sup>c)</sup>

*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138*

(Dated: July 4, 2011)

Vocal tract resonance characteristics in acoustic speech signals are classically tracked using frame-by-frame point estimates of formant frequencies followed by candidate selection and smoothing using dynamic programming methods that minimize *ad hoc* cost functions. The goal of the current work is to provide both point estimates and associated uncertainties of center frequencies and bandwidths in a statistically principled state-space framework. Extended Kalman (K) algorithms take advantage of a linearized mapping to infer formant and antiformant parameters from frame-based estimates of autoregressive moving average (ARMA) cepstral coefficients. Error analysis of KARMA, WaveSurfer, and Praat is accomplished in the all-pole case using a manually marked formant database and synthesized speech waveforms. KARMA formant tracks exhibit lower overall root-mean-square error relative to the two benchmark algorithms, with third formant tracking more challenging. Antiformant tracking performance of KARMA is illustrated using synthesized and spoken nasal phonemes. The simultaneous tracking of uncertainty levels enables practitioners to recognize time-varying confidence in parameters of interest and adjust algorithmic settings accordingly.

PACS numbers: 43.72.Ar, 43.70.Bk, 43.60.Cg, 43.60.Uv

## I. INTRODUCTION

Speech formant tracking has received continued attention over the past sixty years to better characterize formant motion during vowels as well as vowel-consonant boundaries. The de facto approach to resonance estimation involves waveform segmentation and the assumption of an all-pole model characterized by second-order digital resonators (Schafer and Rabiner, 1970). The center frequency and bandwidth of each resonator are then estimated through picking peaks in the all-pole spectrum or finding roots of the prediction polynomial. Tracking these estimates across frames is typically accomplished via dynamic programming methods that minimize cost functions to produce smoothly-varying trajectories.

This general formant-tracking algorithm is implemented in WaveSurfer (Sjölander and Beskow, 2005) and Praat (Boersma and Weenink, 2009), speech analysis tools that enjoy widespread use in the speech recognition, clinical, and linguistic communities. There are, however, numerous shortcomings to this classical approach. For example, formant track smoothing and correction (e.g., for large frequency jumps) are performed in an *ad hoc* manner that precludes that ability to apply statistical

analysis to obtain confidence intervals around the estimated tracks.

Initial development of the formant tracking approach described here has been reported by Rudoy *et al.* (2007) using a manually marked formant database for error analysis (Deng *et al.*, 2006b). The current work continues this line of research and offers two main contributions. The first provides improvements to the Kalman-based autoregressive approach of Deng *et al.* (2007) and extensions to enable antiformant frequency and bandwidth tracking in a Kalman-based autoregressive moving average (KARMA) framework. The second empirically determines the performance of the KARMA approach through visual and quantitative error analysis and compares this performance with that of WaveSurfer and Praat.

### A. Classical formant tracking algorithms

Linear predictive coding (LPC) models have been shown to efficiently encode source/filter characteristics of the acoustic speech signal (Atal and Hanauer, 1971). To extract frame-by-frame formant parameters, the poles of the LPC spectrum can be computed as the roots of the prediction polynomial, peaks in the LPC spectrum, or peaks in the second derivative of the frequency spectrum (Christensen *et al.*, 1976). The first complete formant *tracker* over multiple continuous speech frames incorporated spectral peak-picking, selection of formants from the candidate peaks using continuity constraints, and voicing detection to handle silent and unvoiced speech segments (McCandless, 1974). Extensions to LPC analysis incorporate autoregressive moving average (ARMA) models that added estimates of candidate zeros associated with anti-resonances during consonantal and nasal speech sounds (Steiglitz, 1977; Atal and Schroeder, 1978).

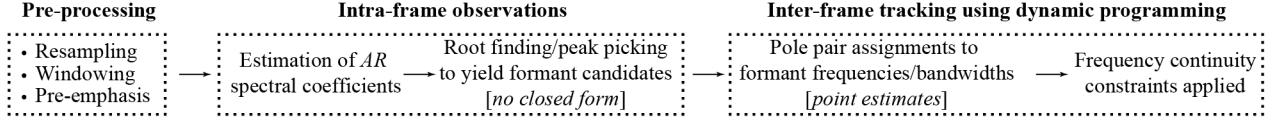
<sup>a)</sup>Portions of this work were presented at the INTERSPEECH conference in Antwerp, Belgium, in August 2007 (Rudoy *et al.*, 2007).

<sup>b)</sup>Electronic address: dmehta@seas.harvard.edu; also with the Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts 02114

<sup>c)</sup>Also with the Department of Statistics, Harvard University, and the Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

A

## CLASSICAL AR APPROACH



B

## PROPOSED KARMA APPROACH

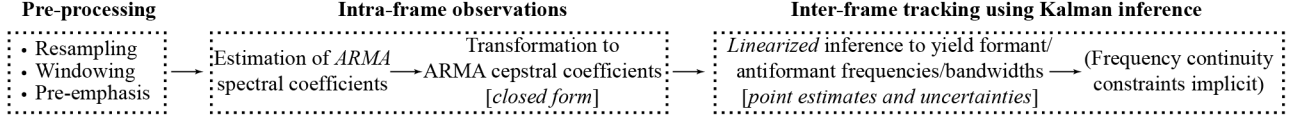


FIG. 1. Illustration of the (A) classical and (B) proposed approaches to formant tracking. Key advantages to the proposed KARMA approach include intra-frame observation of autoregressive moving average (ARMA) parameters for both formant and antiformant tracking, inter-frame tracking using linearized Kalman (K) inference, and the availability of both point estimates and uncertainties for each trajectory.

Fig. 1A illustrates the classical tracking process for the all-pole case. Following pre-processing steps, LPC spectral coefficients yield intra-frame point estimates of candidate frequency and bandwidth parameters via root finding or peak-picking. Intra-frame parameter estimation can be accomplished using a number of methods (Atal and Hanauer, 1971; Atal and Schroeder, 1978; Broad and Clermont, 1989; Yegnanarayana, 1978), and inter-frame parameter selection and smoothing can be performed by minimizing various cost functions in a dynamic programming environment (Sjölander and Beskow, 2005; Boersma and Weenink, 2009). Note that the required root-finding (or peak-picking) procedure cannot be written in closed form. Consequently, statistical analysis (distributions, bias, variance) of the resultant formant and bandwidth estimates is challenging. Alternative spectrographic representations primarily apply to sustained vowels and require significant manual interaction (Fulop, 2010).

## B. Statistical formant tracking algorithms

Probabilistic and statistical models for tracking formants have gained widespread use in the past 15 years with motivation from automatic speech recognition applications. The first such probabilistic model was introduced by Kopec (1986), in which a hidden Markov model was used to constrain the evolution of vector-quantized sets of formant frequencies and bandwidths. Similarly, a state-space dynamical model can appropriately constrain the evolution of formant parameters, where observations of the acoustic speech waveform are linked through non-linear relationships to “hidden” states (formant parameters) that evolve over time. Inference of the state values can be performed by variants of Kalman filter algorithms (Kalman, 1960).

In these algorithms, *ad hoc* assignment of poles and zeros to appropriate formant indices is precluded by the inherent association of spectral/cepstral coefficients to formant and antiformant frequencies and bandwidths.

The first reported state-space approach to formant tracking inferred formant frequencies and bandwidths directly from LPC spectral coefficients (Rigoll, 1986). An extension to this LPC approach was made by Toyoshima *et al.* (1991) to build a tracker that inferred frequencies and bandwidths of both formants and antiformants from time-varying ARMA spectral coefficients (Miyanaga *et al.*, 1986). More recent state-space models define the observations as coefficients in the LPC *cepstral* domain (Zheng and Hasegawa-Johnson, 2004; Deng *et al.*, 2007), providing statistical methods to support or refute empirical relations obtained between low-order cepstral coefficients and formant frequencies (Broad and Clermont, 1989).

The proposed KARMA (Kalman-based autoregressive moving average) approach explores the performance of such a state-space model with ARMA cepstral coefficients as observations to track formant and antiformant parameters. Taking advantage of a linearized mapping between frequency and bandwidth values and cepstral coefficients, KARMA applies Kalman inference to yield point estimates and uncertainties for the output trajectories.

## II. METHODS

Fig. 1B illustrates the proposed statistical modeling approach to formant and antiformant tracking. This approach affords several advantages over classical approaches: (1) both formant and antiformant trajectories are tracked, (2) both frequency and bandwidth estimates are propagated as distributions instead of point estimates to provide for uncertainty quantification, and (3) pole/zero assignment to formants/antiformants is made through a linearized cepstral mapping instead of candidate selection using *ad hoc* cost functions.

## A. Step 1: Pre-processing

The sampled acoustic speech waveform  $s[m]$  is first windowed into short-time frames  $s_t[m] = s[m]w_t[m]$  using overlapping windows  $w_t[m]$  with frame index  $t$ . Each short-time frame  $s_t[m]$  is then pre-emphasized via

$$s_t[m] = s_t[m] - \gamma s_t[m-1], \quad (1)$$

where  $\gamma$  is the pre-emphasis coefficient defining the inherent high-pass filter characteristic that is typically applied to equalize energy across the speech spectrum for improved model fitting.

## B. Step 2: Intra-frame observation generation

### 1. ARMA model of speech

Following windowing and pre-emphasis, the acoustic waveform  $s_t[m]$  is modeled as a stochastic ARMA( $p, q$ ) process:

$$s_t[m] = \sum_{i=1}^p a_i s_t[m-i] + \sum_{j=1}^q b_j u[m-j] + u[m], \quad (2)$$

where  $a_i$  are the  $p$  AR coefficients,  $b_j$  are the  $q$  MA coefficients, and  $u[m]$  is the stochastic excitation waveform. The  $z$ -domain transfer function associated with Eq. (2) is

$$T(z) \triangleq \frac{1 + \sum_{j=1}^q b_j z^{-j}}{1 - \sum_{i=1}^p a_i z^{-i}}. \quad (3)$$

A number of standard spectral estimation techniques can be employed in order to fit data to the ARMA( $p, q$ ) model (see Marelli and Balazs, 2010, for a recent review of ARMA estimation methods). In the current study, ARMA estimation was performed using the ‘armax’ function in MATLAB’s System Identification toolbox (The MathWorks, Natick, MA), which implements an iterative method to minimize a quadratic error prediction criterion (Ljung, 1999, Section 10.2).

### 2. Generation of observations: ARMA cepstral coefficients

In the proposed approach, the ARMA spectral coefficients in Eq. (3) are transformed to the complex cepstrum before inferring formant characteristics. This mapping from ARMA spectral coefficients to ARMA cepstral coefficients has been derived in the all-pole case (e.g., Deng *et al.*, 2006a) and can be extended to account for the presence of zeros in the spectrum. Letting  $C_n$  denote the  $n$ th cepstral coefficient,

$$C_n = c_n - c'_n, \quad (4)$$

where  $C_n$  depends on separate contributions from the denominator and numerator of the ARMA model through the following recursive relationships:

$$c_n = \begin{cases} a_n & \text{if } n = 1 \\ a_n + \sum_{i=1}^{n-1} \binom{i}{n} a_{n-i} c_i & \text{if } 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \binom{i}{n} a_{n-i} c_i & \text{if } p < n, \end{cases} \quad (5a)$$

$$c'_n = \begin{cases} b_n & \text{if } n = 1 \\ b_n + \sum_{j=1}^{n-1} \binom{j}{n} b_{n-j} c'_j & \text{if } 1 < n \leq q \\ \sum_{j=n-q}^{n-1} \binom{j}{n} b_{n-j} c'_j & \text{if } q < n. \end{cases} \quad (5b)$$

Derivation of Eqs. (5) is given in the Appendix. The proof is derived under the minimum-phase assumption that constrains the poles and zeros of the ARMA transfer function to lie within the unit circle.

## C. Step 3: Inter-frame parameter tracking

The proposed algorithm tracks point estimates and uncertainties for  $I$  formants and  $J$  antiformants from frame to frame. To accommodate the temporal dimension, the parameters of frame  $t$  are placed in column vector  $\mathbf{x}_t$ :

$$\mathbf{x}_t \triangleq (f_1 \dots f_I \ b_1 \dots b_I \ f'_1 \dots f'_J \ b'_1 \dots b'_J)^T, \quad (6)$$

where  $(f_i, b_i)$  is the frequency/bandwidth pair of the  $i$ th formant and  $(f'_j, b'_j)$  is the frequency/bandwidth pair for the  $j$ th antiformant.

### 1. Observation model

Inference of the output parameters is facilitated by a closed-form mapping from the state vector  $\mathbf{x}_t$  to the observed cepstral coefficients  $C_n$  in Eq. (4). Extending the speech production model of Schafer and Rabiner (1970) to capture zeros, we assume that the transfer function  $T(z)$  of the ARMA model can be written as a cascade of  $I$  second-order digital resonators and  $J$  second-order digital anti-resonators:

$$T(z) = \frac{\prod_{j=1}^J (1 - \beta_j z^{-1})(1 - \bar{\beta}_j z^{-1})}{\prod_{i=1}^I (1 - \alpha_i z^{-1})(1 - \bar{\alpha}_i z^{-1})}, \quad (7)$$

where  $(\alpha_i, \bar{\alpha}_i)$  and  $(\beta_j, \bar{\beta}_j)$  denote complex-conjugate pole and zero pairs, respectively. Each pole and zero are parameterized by a center frequency and 3-dB bandwidth (both in units of Hertz) using the following relations:

$$(\alpha_i, \bar{\alpha}_i) = \exp\left(\frac{-\pi b_i \pm 2\pi\sqrt{-1}f_i}{f_s}\right), \quad (8a)$$

$$(\beta_j, \bar{\beta}_j) = \exp\left(\frac{-\pi b'_j \pm 2\pi\sqrt{-1}f'_j}{f_s}\right), \quad (8b)$$

where  $f_s$  is the sampling rate (in Hz).

Performing a Taylor-series expansion of  $\log T(z)$  yields

$$\log T(z) = \sum_{i=1}^I \sum_{n=1}^{\infty} \frac{(\alpha_i^n + \bar{\alpha}_i^n)}{n} z^{-n} - \sum_{j=1}^J \sum_{n=1}^{\infty} \frac{(\beta_j^n + \bar{\beta}_j^n)}{n} z^{-n}. \quad (9)$$

Recalling that  $C_n$  is the  $n$ th cepstral coefficient,  $\log T(z) = C_0 + \sum_{n=1}^{\infty} C_n z^{-n}$ . Thus, equating the coefficients of powers of  $z^{-1}$  leads to

$$C_n = \frac{1}{n} \sum_{i=1}^I (\alpha_i^n + \overline{\alpha_i^n}) - \frac{1}{n} \sum_{j=1}^J (\beta_j^n + \overline{\beta_j^n}). \quad (10)$$

Finally, inserting  $\alpha_i$  and  $\beta_j$  from Eqs. (8) into Eq. (10) yields the following observation model  $h(\mathbf{x}_t)$  that maps elements of  $\mathbf{x}_t$  to  $C_n$ :

$$\begin{aligned} h(\mathbf{x}_t) &\triangleq C_n \\ &= \frac{2}{n} \sum_{i=1}^I \exp\left(-\frac{\pi n}{f_s} b_i\right) \cos\left(\frac{2\pi n}{f_s} f_i\right) - \\ &\quad \frac{2}{n} \sum_{j=1}^J \exp\left(-\frac{\pi n}{f_s} b'_j\right) \cos\left(\frac{2\pi n}{f_s} f'_j\right). \end{aligned} \quad (11)$$

## 2. State-space model

We adopt a state-space framework similar to that by Deng *et al.* (2007) to model the evolution of the state vector in Eq. (6) from frame  $t$  to frame  $t+1$ :

$$\mathbf{x}_{t+1} = \mathbf{F} \mathbf{x}_t + \mathbf{w}_t, \quad (12a)$$

$$\mathbf{y}_t = h(\mathbf{x}_t) + \mathbf{v}_t, \quad (12b)$$

where  $\mathbf{F}$  is the state transition matrix, and  $\mathbf{w}_t$  and  $\mathbf{v}_t$  are uncorrelated white Gaussian sequences with covariance matrices  $\mathbf{Q}$  and  $\mathbf{R}$ , respectively. The function  $h(\mathbf{x}_t)$  is the nonlinear mapping of Eq. (11), and vector  $\mathbf{y}_t$  consists of *estimates* of the first  $N$  cepstral coefficients of  $C_n$  (not including the zeroth coefficient). The initial state  $\mathbf{x}_0$  follows a normal distribution with mean  $\boldsymbol{\mu}_0$  and covariance  $\boldsymbol{\Sigma}_0$ . The state-space model of Eqs. (12) is thus parameterized by the set  $\boldsymbol{\theta}$ :

$$\boldsymbol{\theta} \triangleq (\mathbf{F}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \quad (13)$$

## 3. Linearization via Taylor approximation

The cepstral mapping in Eq. (11) can be linearized to enable *approximate* minimum-mean-square-error (MMSE) estimates of the tracked states via the extended Kalman filter. The mapping  $h(\mathbf{x}_t)$  is linearized by computing the first-order terms of the Taylor-series expansion of  $C_n$  in Eq. (11):

$$\begin{aligned} \frac{\partial C_n}{\partial f_i} &= -\frac{4\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b_i\right) \sin\left(\frac{2\pi n}{f_s} f_i\right), \\ \frac{\partial C_n}{\partial b_i} &= -\frac{2\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b_i\right) \cos\left(\frac{2\pi n}{f_s} f_i\right), \\ \frac{\partial C_n}{\partial f'_j} &= \frac{4\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b'_j\right) \sin\left(\frac{2\pi n}{f_s} f'_j\right), \\ \frac{\partial C_n}{\partial b'_j} &= \frac{2\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b'_j\right) \cos\left(\frac{2\pi n}{f_s} f'_j\right). \end{aligned}$$

TABLE I. Extended Kalman smoother algorithm.

1. Initialization: Set  $\mathbf{m}_{0|0} = \boldsymbol{\mu}_0$  and  $\mathbf{P}_{0|0} = \boldsymbol{\Sigma}_0$
2. Filtering: Repeat for  $t = 1, \dots, T$

$$\begin{aligned} \mathbf{m}_{t|t-1} &= \mathbf{F} \mathbf{m}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{F} \mathbf{P}_{t-1|t-1} \mathbf{F}^T + \mathbf{Q} \\ \mathbf{K}_t &= \mathbf{P}_{t|t-1} \mathbf{H}_t^T (\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R})^{-1} \\ \mathbf{m}_{t|t} &= \mathbf{m}_{t|t-1} + \mathbf{K}_t (\mathbf{y}_t - h(\mathbf{m}_{t|t-1})) \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_{t|t-1} \end{aligned} \quad (16)$$

3. Smoothing: Repeat for  $t = T, \dots, 1$

$$\begin{aligned} \mathbf{S}_t &= \mathbf{P}_{t-1|t-1} \mathbf{F}^T \mathbf{P}_{t|t-1}^{-1} \\ \mathbf{m}_{t-1|T} &= \mathbf{m}_{t-1|t-1} + \mathbf{S}_t (\mathbf{m}_{t|T} - \mathbf{F} \mathbf{m}_{t-1|t-1}) \\ \mathbf{P}_{t-1|T} &= \mathbf{P}_{t-1|t-1} + \mathbf{S}_t (\mathbf{P}_{t|T} - \mathbf{P}_{t-1|t-1}) \mathbf{S}_t^T \end{aligned}$$

The Jacobian matrix  $\mathbf{H}_t$  thus consists of four sub-matrices for each frame  $t$ :

$$\mathbf{H}_t \triangleq (\mathbf{H}(f_i) \quad \mathbf{H}(b_i) \quad \mathbf{H}(f'_j) \quad \mathbf{H}(b'_j)), \quad (14)$$

where  $\mathbf{H}(f_i)$  and  $\mathbf{H}(b_i)$  each consists of  $N$  rows and  $p/2$  columns:

$$\mathbf{H}(f_i) \triangleq \begin{pmatrix} \frac{\partial C_1}{\partial f_1} & \frac{\partial C_1}{\partial f_2} & \dots & \frac{\partial C_1}{\partial f_{p/2}} \\ \frac{\partial C_2}{\partial f_1} & \frac{\partial C_2}{\partial f_2} & \dots & \frac{\partial C_2}{\partial f_{p/2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial C_N}{\partial f_1} & \frac{\partial C_N}{\partial f_2} & \dots & \frac{\partial C_N}{\partial f_{p/2}} \end{pmatrix}, \quad (15a)$$

$$\mathbf{H}(b_i) \triangleq \begin{pmatrix} \frac{\partial C_1}{\partial b_1} & \frac{\partial C_1}{\partial b_2} & \dots & \frac{\partial C_1}{\partial b_{p/2}} \\ \frac{\partial C_2}{\partial b_1} & \frac{\partial C_2}{\partial b_2} & \dots & \frac{\partial C_2}{\partial b_{p/2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial C_N}{\partial b_1} & \frac{\partial C_N}{\partial b_2} & \dots & \frac{\partial C_N}{\partial b_{p/2}} \end{pmatrix}, \quad (15b)$$

and  $\mathbf{H}(f'_j)$  and  $\mathbf{H}(b'_j)$  are defined analogously each with  $N$  rows and  $q/2$  columns.

## 4. Kalman-based inference

Given observations  $\mathbf{y}_t$  for frame indices 1 to  $T$ , the extended Kalman smoother (EKS) can be used to compute the mean  $\mathbf{m}_{t|T}$  (point estimates) and covariance  $\mathbf{P}_{t|T}$  (estimate uncertainties) of each parameter in  $\mathbf{x}_t$ . Table I displays the steps of the EKS, which employs a two-pass filtering (forward) and smoothing (backward) procedure. For real-time processing, the forward filtering stage may be applied without a backward smoothing procedure; naturally, this will lead to larger uncertainties in the corresponding parameter estimates.

Care must be taken when approximating the observation model of Eq. (11) to avoid suboptimal performance or algorithm divergence in the case of the Kalman filter (Julier and Uhlmann, 1997). To verify the appropriateness of the linearization in Section II.C.3 in this

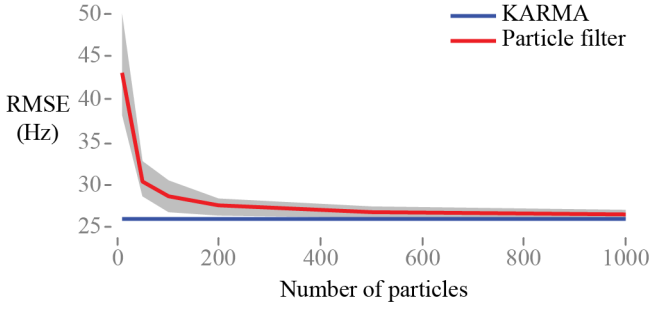


FIG. 2. (Color online) Comparison of KARMA (blue) and particle filter (red) tracking performance in terms of root-mean-square error (RMSE) averaged over 25 Monte Carlo trials and reported with 95 % confidence intervals (gray).

setting, comparisons are made to a more computationally intensive method of stochastic computation termed a particle filter, which approximates the densities in question by sequentially propagating a fixed number of samples, or “particles,” and hence avoids the linearization of Eq. (11).

Twenty-five Monte Carlo simulations of 100-sample data sequences were performed according to Eqs. (12) with four complex-conjugate pole pairs ( $p = 8$ ,  $q = 0$ ) and  $N = 15$  cepstral coefficients. Fig. 2 compares the output of KARMA using the extended Kalman filter of Table I and that of a particle filter in terms of root-mean-square error (RMSE) as a function of the number of particles, averaged over all formant frequency tracks (true bandwidths were provided to both algorithms). The performance of the EKF compares favorably to that of the particle filter, even when a large number of particles is used. Similar results hold over a broad range of parameter values.

### 5. Observability of states

The model of Eq. (12) does not explicitly take into account the existence of speech and non-speech states. To continue to track or *coast* parameters during silence frames, the state vector  $\mathbf{x}_t$  can be augmented with a binary indicator variable to specify the presence of speech in the frame. The approximate MMSE state estimate is then obtained via EKS inference by modifying the Kalman gain in Eq. 16:

$$\mathbf{K}_t = \mathbf{M}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T (\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R})^{-1},$$

where  $\mathbf{M}_t$  is a diagonal matrix with diagonal entries equal to 1 or 0 depending on the presence or absence, respectively, of speech energy in frame  $t$ .

In addition, to handle the presence or absence of particular tracks, the state vector  $\mathbf{x}_t$  can be dynamically modified to include or omit corresponding frequency/bandwidth states in Eq. (6). The approximate MMSE state estimate is then obtained via EKS inference in Table I with the modified state vector. If an absent

state reappears in a given frame, that state is reinitialized with corresponding entries in  $\boldsymbol{\mu}_0$  and  $\boldsymbol{\Sigma}_0$ .

### 6. Model order selection and system identification

As is commonly done, the orders  $p$  and  $q$  of the ARMA model are chosen to capture as much information as possible on the peaks and valleys in the resonance spectrum, while avoiding overfitting and mistakenly capturing source-related information. The ARMA cepstral order  $N$  is chosen to be at least  $\max(p, q)$  so that all pole/zero information is incorporated per Eq. (5). Finally, selecting  $I$  and  $J$  in Eq. (11) depends on the expected number of formants and anti-formants, respectively, in the speech bandwidth  $f_s/2$ .

Formants do not evolve independently of one another, and their temporal trajectories are not independent in frequency. In the synthesis of front vowels, it is common practice to employ a *linear regression* of  $f_3$  onto  $f_1$  and  $f_2$  (Nearey, 1989, e.g.). Empirically, we found the formant cross-correlation function to decay slowly (Rudoy *et al.*, 2007), implying that a set of formant values at frame  $t$  might be helpful in predicting values of all formants at frame  $t + 1$ . Thus, instead of setting the state transition matrix  $\mathbf{F}$  to the identity matrix (Deng *et al.*, 2006a, 2007),  $\mathbf{F}$  is estimated *a priori* for a particular utterance from first-pass WaveSurfer formant frequency tracks using a linear least-squares estimator (Hamilton, 1994).

The state transition covariance matrix  $\mathbf{Q}$ , which dictates the frame-to-frame frequency variation, consists of a diagonal matrix with values corresponding to standard deviations of approximately 320 Hz for center frequencies and 100 Hz for bandwidths. These values were empirically found to follow temporal variations of speech articulation. The covariance matrix  $\mathbf{R}$ , representing the signal-to-noise ratio of the cepstral coefficients, is a diagonal matrix with elements  $\mathbf{R}_{nn} = 1/n$  for  $n \in \{1, 2, \dots, N\}$ . This was observed to be in reasonable agreement with the variance of the residual vector of the cepstral coefficients derived from speech waveforms.

The center frequencies and bandwidths are initialized to  $\boldsymbol{\mu}_0 = (500 \ 1500 \ 2500 \ 80 \ 120 \ 160)$  Hz. The initial covariance  $\boldsymbol{\Sigma}_0$  is set to  $\mathbf{Q}$ .

### D. Summary of KARMA approach

Table II outlines the steps of the proposed KARMA algorithm, which includes a pre-processing stage, intra-frame ARMA cepstral coefficient estimation, and inter-frame tracking of formant and antiformant parameters using Kalman inference.

### E. Benchmark algorithms

Performance of KARMA is compared with that of Praat (Boersma and Weenink, 2009) and WaveSurfer (Sjölander and Beskow, 2005), two software packages

TABLE II. Proposed KARMA algorithm for formant and antiformant tracking.

---



---

Repeat for frames $t = 1, \dots, T$ (Online or batch mode)	
1. Pre-processing of input speech waveform $s[m]$	
(a) Window: $s_t[m] = s[m]w_t[m]$	
(b) Pre-emphasize $s_t[m]$	
2. Intra-frame observation of $N$ cepstral coefficients	
(a) Estimate ARMA( $p, q$ ) spectral coefficients $\hat{a}_i$ and $\hat{b}_j$ in Eq. (3)	
(b) Convert $\hat{a}_i$ and $\hat{b}_j$ to ARMA cepstral coefficients using Eq. (4)	
3. Inter-frame parameter tracking of $I$ formants and $J$ antiformants	
(a) Apply Kalman filtering step in Table I	
(b) $\mathbf{m}_{t t}$ are point estimates and diagonal elements of $\mathbf{P}_{t t}$ are associated variances of the estimates	
Repeat for frames $t = T, \dots, 1$ (Batch mode only)	
4. Inter-frame parameter tracking of $I$ formants and $J$ antiformants	
(a) Apply Kalman smoothing step in Table I	
(b) $\mathbf{m}_{t T}$ are point estimates and diagonal elements of $\mathbf{P}_{t T}$ are associated variances of the estimates	

---



---

that see wide use among voice and speech researchers. WaveSurfer and Praat both follow the classical formant tracking approach in which frame-by-frame formant frequency candidates are obtained from the all-pole spectrum and smoothed across the entire speech utterance to remove outliers and constrain the trajectories to physiologically plausible values. Smoothing is accomplished through dynamic programming to minimize the sum of the following three cost functions: (1) the deviation between the frequency for each formant from baseline values of each frequency; (2) a measure of the quality factor  $f_i/b_i$  of a formant, where higher quality factors are favored; and (3) a transition cost that penalizes large frequency jumps. The user sets weights to these cost functions to tune the algorithm's performance.

### III. RESULTS

Evaluation of KARMA is accomplished in the all-pole case using the vocal tract resonance (VTR) database (Deng *et al.*, 2006b). Since the VTR database itself only yields estimates of ground truth and exhibits observable labeling errors, two speech databases are created using overlap-add of synthesis speech frames using the four VTR formant tracks. Antiformant tracking performance of KARMA is illustrated using synthesized and spoken nasal phonemes.

#### A. Error analysis using a hand-corrected formant database

The VTR database contains a representative subset of the TIMIT speech corpus (Garofolo *et al.*, 1993) that consists of 516 diverse, phonetically-balanced utterances collated across gender, individual speakers, dialects, and phonetic contexts. The VTR database contains state information for four formant trajectory pairs (center frequency and bandwidth). The first *three* center frequency trajectories were manually corrected after an initial automated pass (Deng *et al.*, 2004). Corrections were made using knowledge-based intervention based on the speech waveform, its wideband spectrogram, word- and phoneme-level transcriptions, and phonemic boundaries.

Analysis parameters of KARMA are set to the following values:  $f_s = 7$  kHz, 20 ms Hamming windows with 50 % overlap,  $\gamma = 0.7$ ,  $p = 12$  ( $q = 0$ ), and  $I = 3$  ( $J = 0$ ). Each frame is fit with an ARMA(12,0) model using the autocorrelation method of linear prediction and, subsequently, transformed to  $N = 15$  cepstral coefficients via Eq. (5). The initial state vector is set to  $\mathbf{x}_0 = (500 \ 1500 \ 2500)^T$ , and  $\Sigma_0$  is set to  $\mathbf{Q}$ . TIMIT phone transcriptions are used to indicate whether each frame contains speech energy or a silence region. A frame is considered silent if all its samples are labeled as a pause (*pau*, *epi*, *h#*), closure interval (*bcl*, *dcl*, *gcl*, *pcl*, *tcl*, *kcl*), or glottal stop (*q*). Thus, errors due to speech activity detection are minimized, and all tracks are coasted during silent frames.

Default smoothing settings are set within WaveSurfer and Praat. Other analysis parameters are matched to KARMA:  $f_s = 7$  kHz, 20 ms Hamming windows with 50 % overlap,  $\gamma = 0.7$ ,  $p = 12$  ( $q = 0$ ), and  $I = 3$  ( $J = 0$ ).

Table III summarizes the performance of KARMA, WaveSurfer, and Praat on the VTR database. The root-mean-square error (RMSE) per formant is computed over all speech frames for each utterance and then averaged, per formant, across all 516 utterances in the database. The cepstral-based KARMA approach results in lower overall error compared to the classical algorithms, with particular gains for  $f_1$  and  $f_2$  tracking. Praat exhibits the lowest average error for  $f_3$ .

Figure 3 illustrates the formant tracks output from the three algorithms for VTR utterance 200 spoken by an adult female. During non-speech regions (the first 700 ms exhibits noise energy during inhalation), mean KARMA

TABLE III. Formant tracking performance of KARMA, WaveSurfer, and Praat in terms of root-mean-square error (RMSE) per formant averaged across 516 utterances in the VTR database (Deng *et al.*, 2006b). RMSE is only computed over speech-labeled frames.

---



---

Formant	KARMA	WaveSurfer	Praat
$f_1$	114 Hz	170 Hz	185 Hz
$f_2$	226 Hz	276 Hz	254 Hz
$f_3$	320 Hz	383 Hz	303 Hz
Overall	220 Hz	276 Hz	247 Hz

---



---

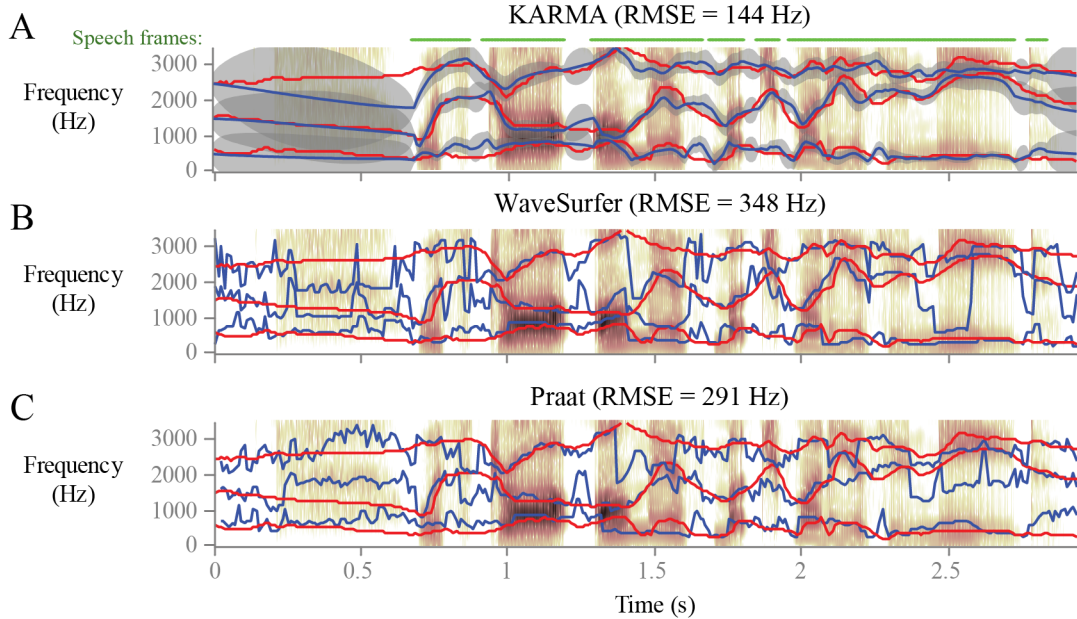


FIG. 3. (Color online) Estimated formant tracks on spectrogram of VTR utterance 200: “Withdraw only as much money as you need.” Reference trajectories from the VTR database are shown in red along with the formant frequency tracks in blue from (A) KARMA, (B) WaveSurfer, and (C) Praat. The KARMA output additionally displays uncertainties (gray shading,  $\pm 1$  standard deviation) for each formant trajectory and speech-labeled frames (green). Reported root-mean-square error (RMSE) is averaged across formants conditioned on speech presence for each frame.

track estimates are linear with increasing uncertainty for frames that are farther from frames with formant information. Compared to the WaveSurfer and Praat tracks, KARMA trajectories are smoother and better behaved, reflecting the slow-moving nature of the speech articulators. The classical algorithms exhibit errant tracking of  $f_2$  during the /i/ vowel in “need” at 2.5 s that is handled by the KARMA approach.

### B. Error analysis using synthesized databases

Speech waveforms in the first database (VTRsynth) are synthesized through overlap-add of frames that each follow the ARMA model of Eq. (2). ARMA spectral coefficients are derived from the four formant frequency/bandwidth pairs in the corresponding frame of the VTR database utterance using the impulse-invariant transformation of a digital resonator (Klatt, 1980). The source excitation is white Gaussian noise during non-silence frames. Synthesis parameters are set to the following values:  $f_s = 16$  kHz, 20 ms Hanning windows with 50 % overlap, and  $p = 8$  ( $q = 0$ ).

The second database (VTRsynthf0) introduces a model mismatch between synthesis and KARMA analysis by applying a Rosenberg C source waveform (Rosenberg, 1971) instead of white noise for each frame considered voiced. The fundamental frequency of each voiced frame in the original VTR database is estimated by WaveSurfer. The VTRsynthf0 database thus includes voiced, unvoiced, and non-speech frames. Synthesis parameters are set as in the VTRsynth database. Formant

trajectories from these two synthesized databases act as truer ground truth contours than in the VTR database to test the performance of the cepstral-based KARMA algorithm.

Table IV and Table V display performance on the VTRsynth and VTRsynthf0 databases, respectively, of the three tested algorithms with settings as described in the previous section. The proposed KARMA approach compares favorably to WaveSurfer and Praat. The similar error of KARMA and WaveSurfer validates the use of ARMA cepstral coefficients as observations in place of ARMA spectral coefficients.

### C. Antiformant tracking

The KARMA approach to formant and antiformant tracking is illustrated in this section. Synthesized and real speech examples are presented to determine the abil-

TABLE IV. Average RMSE of KARMA, WaveSurfer, and Praat formant tracking of the first three formant trajectories in the VTRsynth database that resynthesizes utterances using a stochastic source and formant tracks from the VTR database. Error is only computed over speech-labeled frames.

Formant	KARMA	WaveSurfer	Praat
$f_1$	29 Hz	37 Hz	58 Hz
$f_2$	53 Hz	60 Hz	123 Hz
$f_3$	64 Hz	54 Hz	130 Hz
Overall	48 Hz	50 Hz	104 Hz



ity of the ARMA-derived cepstral coefficients to capture pole and zero information.

### 1. Synthesized waveform

In the synthesized case, a speech-like waveform /nan/ is generated with varying frame-by-frame formant and antiformant characteristics and a periodic source excitation as was implemented for the VTRsynth0 database. The /nan/ waveform is synthesized at  $f_s = 10$  kHz using 75 100 ms frames with 50 % overlap. Formant frequencies (bandwidths) of the /n/ phonemes are set to 257 Hz (32 Hz) and 1891 Hz (100 Hz). One antiformant is placed at 1223 Hz (bandwidth of 52 Hz) to mimic the location of an alveolar nasal antiformant. Formant frequencies (bandwidths) of /a/ were set to 850 Hz (80 Hz) and 1500 Hz (120 Hz). A random term with zero mean and standard deviation of 10 Hz was added to each trajectory to simulate realistic variation.

Fig. 4 shows the results of formant and antiformant tracking using KARMA on the synthesized phoneme string /nan/. Two different visualizations are displayed. Fig. 4A plots point estimates and uncertainties of the center frequency *and* bandwidth trajectories for each frame. Fig. 4B displays the wideband spectrogram with overlaid center frequency tracks whose width reflects the corresponding 3-dB bandwidth value. Note that the length of the state vector in the KARMA’s state-space model is modified depending on the presence or absence of antiformant energy. Estimated trajectories fit the ground truth values well once initialized values reach a steady state.

### 2. Spoken nasals

During real speech, a vocal tract configuration consisting of multiple acoustic paths results in the possible existence of both poles and zeros in transfer function  $T(z)$  (Eq. 7). For example, the effects of antiresonances might enter the transfer function of nasalized speech sounds as zeros in  $T(z)$ . Typically, the frequency of the lowest zero depends on tongue position. For the labial nasal consonant /m/, the frequency of this antiresonance is approximately 1100 Hz. As the point of closure moves toward the back of the oral cavity—such as for the alveolar and velar

nasal consonants—the length of the resonator decreases, and the frequency of this zero increases. The frequency of a second zero is approximately three times the frequency of the lowest zero due to the quarter-wavelength oral cavity configuration.

KARMA performance was evaluated visually on spoken nasal consonants produced with closure at the labial (/m/), alveolar (/n/), and velar (/ŋ/) positions. The extended Kalman smoother was applied using an ARMA(16, 4) model,  $f_s = 8$  kHz, 20 ms Hamming windows with 50 % overlap,  $N = 20$  cepstral coefficients, and  $\gamma = 0.7$ . The frequencies (bandwidths) of the formants were initialized to 500 Hz (80 Hz), 1500 Hz (120 Hz), and 2500 Hz (160 Hz). The frequencies (bandwidths) of the antiformants were initialized to 1000 Hz (80 Hz) and 2000 Hz (80 Hz).

Figure 5 displays KARMA outputs (point estimate and uncertainty of frequency tracks) and averaged spectra for the three sustained consonants. The KARMA algorithm takes a few frames to settle to its steady-state estimates. As expected, the frequency of the antiformant increases as the position of closure moves toward the back of the oral cavity. The uncertainty of the first antiformant of /ŋ/ increases significantly, indicating that this antiformant is not well observed in the waveform. Note that the inclusion of zeros greatly improves the ability of the ARMA model to fit the underlying waveform spectra.

Finally, the KARMA tracker was applied to the spoken word “piano” to determine if the antiformant tracks would capture any zeros during the nasal phoneme. Figure 6 displays the KARMA formant and antiformant tracks with their associated uncertainties. During the non-nasalized regions, the uncertainty around the point estimates of the antiformant track is large, reflecting the lack of antiresonance information. During the /n/ segment, the uncertainty of the antiformant tracks decreases to reveal observable antiformant information.

## IV. DISCUSSION

In this article, the task of tracking frequencies and bandwidths of formants and antiformants was approached from a statistical point of view. The evolution of parameters was cast in a state-space model to provide access to point estimates and uncertainties of each track. The key relationship was a linearized mapping between cepstral coefficients and formant and antiformant parameters that allowed for the use of the extended family of Kalman inference algorithms.

The VTR database provides an initial benchmark of “ground truth” for the first three formant frequency values to which multiple algorithm outputs can be compared. The values in the VTR database, however, should be interpreted with caution because starting values were initially obtained via a first-pass automatic algorithm Deng *et al.* (2004). It is unclear how much manual intervention was required and what types of errors were corrected. In particular, VTR tracks do not always overlap high-energy spectral regions. Despite the presence of various labeling errors in the VTR database, it is still

TABLE V. Average RMSE of KARMA, WaveSurfer, and Praat formant tracking of the first three formant trajectories in the VTRsynth0 database that resynthesizes VTR database utterances using stochastic and periodic sources. Error is only computed over speech-labeled frames.

Formant	KARMA	WaveSurfer	Praat
$f_1$	44 Hz	57 Hz	57 Hz
$f_2$	53 Hz	58 Hz	117 Hz
$f_3$	62 Hz	59 Hz	111 Hz
Overall	53 Hz	58 Hz	95 Hz



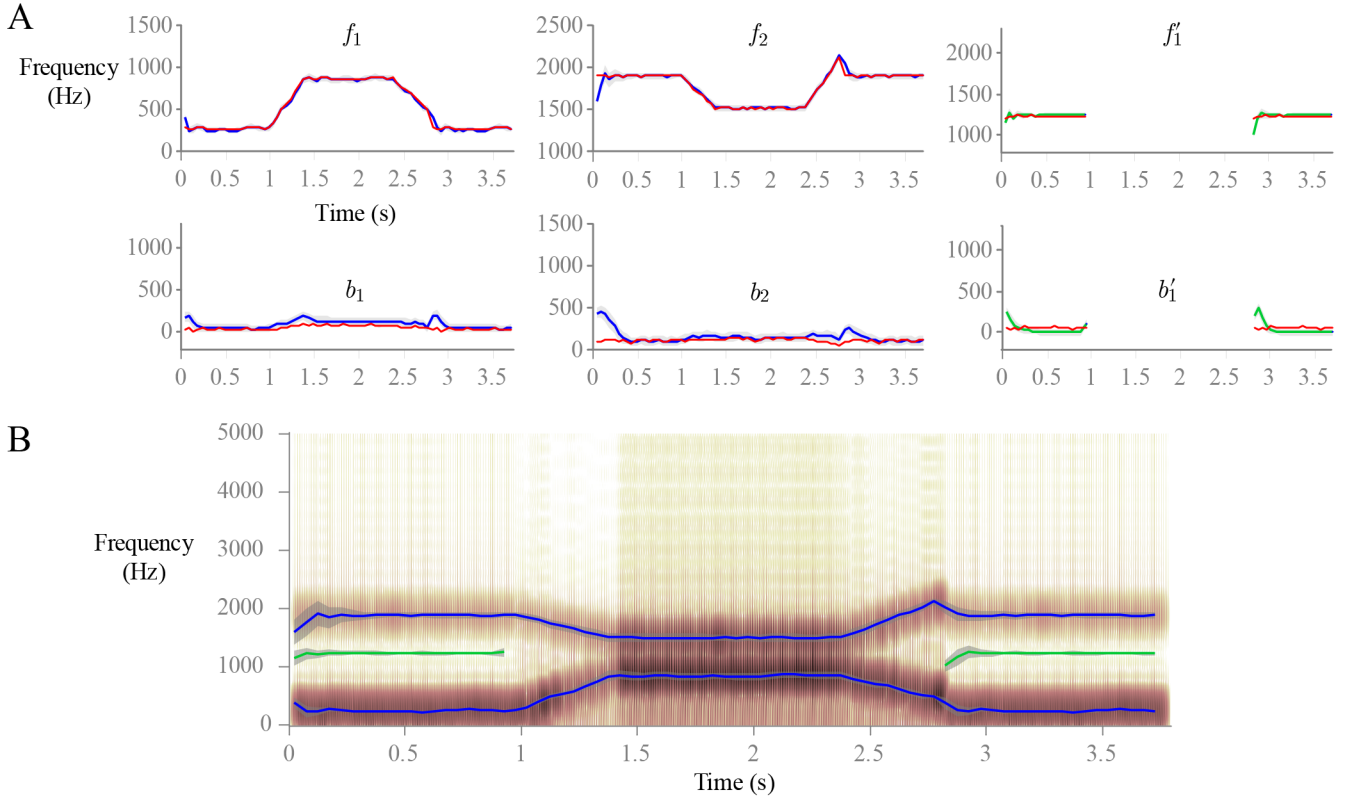


FIG. 4. (Color online) Illustration of the output from KARMA for the synthesized utterance /nan/. Plots in panel A overlay the true trajectories (red) with the mean estimates (blue for formants, green for antiformants) and uncertainties (gray shading) for each frequency and bandwidth. Panel B plots an alternative display with a wideband spectrogram along with estimated frequency and bandwidth tracks of formants (blue) and antiformants (green). The 3-dB bandwidths dictate the width of the corresponding frequency tracks.

useful to obtain initial performance of formant tracking algorithms on real speech.

In the current framework, cepstral coefficients are derived from the spectral coefficients of the fitted stochastic ARMA model (Section II.B). Source information related to phonation is thus separated from vocal tract resonances by assuming that the source is a white Gaussian noise process. This is not the case in reality, especially for voiced speech, where the source excitation component has its own characteristics in frequency (e.g., spectral slope) and time (e.g., periodicity). This model mismatch has been explored here via VTRsynthf0, though we note that it is also possible to incorporate more sophisticated source modeling through the use flexible basis functions such as wavelets (Mehta *et al.*, 2011).

An alternative approach to ARMA modeling is to compute the nonparametric (real) cepstrum directly from the speech samples. Based on the convolutional model of speech, low-frequency cepstral coefficients are largely linked to vocal tract information up to about 5 ms (Childers *et al.*, 1977), depending on the fundamental frequency. Skipping the 0th cepstral coefficient that quantifies the overall spectral level, this would translate to including up to the first 35 cepstral coefficients in the observation vector  $\mathbf{y}_t$  in the state-space framework (Eq. 12). Although coefficients from the real cepstrum do

not strictly adhere to the observation model derived in Eq. 11, the approximate separation of source and filter in the nonparametric cepstral domain makes this approach viable.

Figure 7 illustrates the output of the algorithm using the first 15 coefficients of the real cepstrum as observations. Interestingly the performance of the nonparametric cepstrum is comparably to that of the parametric ARMA cepstrum, especially for the first formant frequency. Most of the error stems from underestimating the second and third formant frequencies. Advantages to using the nonparametric cepstrum include computational efficiency and freedom from ARMA model constraints.

The capability of automated methods to track the third formant strongly depends on the resampling frequency, which controls the amount of energy in the spectrum at higher frequencies. For example, if the signal were resampled to 10 kHz, a given algorithm might erroneously track the third formant frequency through spectral regions typically ascribed to the fourth formant. Traditional formant tracking algorithms have access to multiple candidate frequencies, which are constantly resorted so that  $f_4 > f_3 > f_2 > f_1$ . In the proposed statistical approach, the ordering of formant indices is inherent in the mapping of formants to cepstral coefficients (11), and further empirical study of this formants-to-

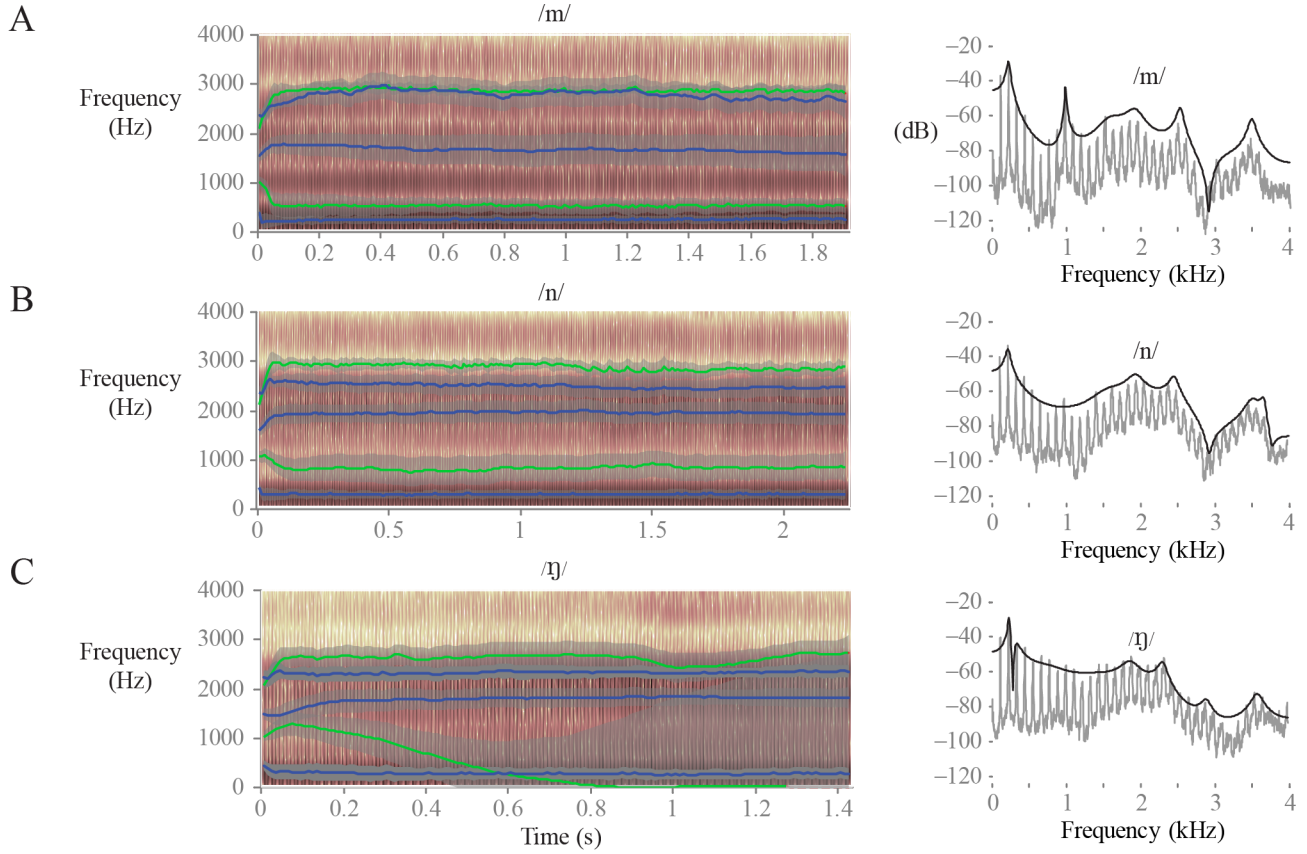


FIG. 5. (Color online) KARMA output for three spoken nasals: (A) /m/, (B) /n/, and (C) /ŋ/. On the left, spectrograms overlay the mean estimates (blue for formants, green for antiformants) and uncertainties (gray shading) for each frequency and bandwidth. Plots to the right display the corresponding periodogram (gray) and spectral ARMA model fit (black).

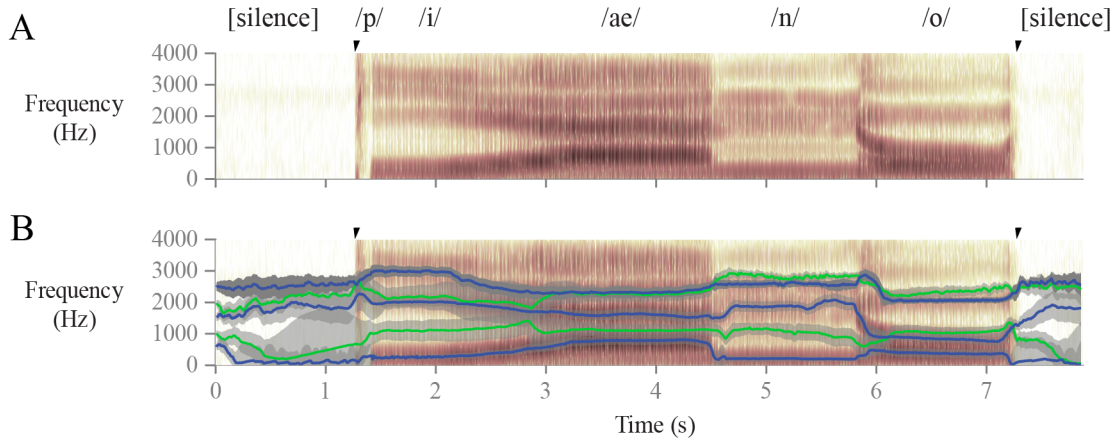


FIG. 6. (Color online) KARMA formant and antiformant tracks of utterance by adult male: “piano.” Displayed are the (A) wideband spectrogram of the speech waveform and (B) the spectrogram overlaid with formant frequency estimates (blue), antiformant frequency estimates (green), and uncertainties ( $\pm 1$  standard deviation) for each track (gray). Arrows indicate beginning and ending of utterance. Note that the increase in uncertainty during silence regions.

cepstrum mapping can be expected to lead to improved methods when there are additional resonances present in the speech bandwidth. Further analysis of “noise” in the estimated ARMA cepstrum can also be expected to improve overall robustness in the presence of various sources

of uncertainty (Tournet and Lacaze, 1995).

Overall, the proposed KARMA approach compares favorably with WaveSurfer and Praat in terms of root-mean-square error. RMSE, however, is only one selected error metric, which must be validated by observing how

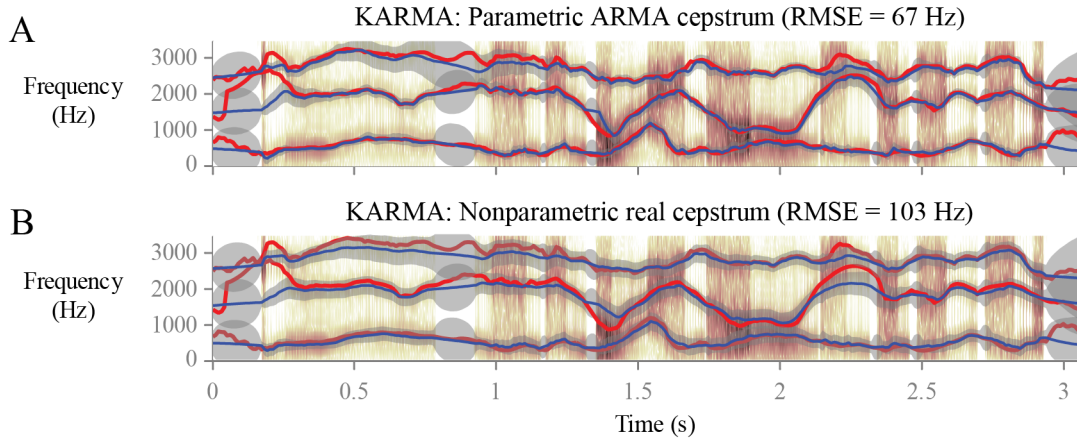


FIG. 7. (Color online) KARMA formant tracks using observations from (A) parametric ARMA cepstrum and (B) nonparametric real cepstrum for VTRsynthf0 utterance 1: “Even then, if she took one step forward, he could catch her.” Reference trajectories from the VTR database are shown in red with the outputs of KARMA in blue. KARMA uncertainties ( $\pm 1$  standard deviation) are shown as gray shading. Reported root-mean-square error (RMSE) averages across 3 formants conditioned on the presence of speech energy for each frame.

well raw trajectories behave. The proposed KARMA tracker yields smoother outputs as a result and offers parameters that allow the user to tune the performance of the algorithm in a statistically principled manner. Such well behaved trajectories may be particularly desirable for the resynthesis of perceptually natural speech.

Though considerably more complex and more sensitive to model assumptions, a time-varying autoregressive moving average (TV-ARMA) model has been previously proposed for formant and antiformant tracking (Toyoshima *et al.*, 1991) with little follow-up investigation. In their study, Toyoshima *et al.* used an extended Kalman filter to solve for ARMA spectral coefficients at each speech sample. One real zero and one real pole were included to model changes in gross spectral shape over time. While a frame-based approach (as taken in KARMA) appears to yield more salient parameters at a lower computational cost, future work could consider this as well as alternative time-varying approaches (Rudoy *et al.*, 2011).

As a final observation, antiformant tracking remains a challenging task in speech analysis. Antiresonances are typically less strong than their resonant counterparts during nasalized phonation, and the estimation of subglottal resonances continues to rely on empirical relationships rather than direct acoustic observation (Arsikere *et al.*, 2011). Nevertheless, the proposed approach allows the user the option of tracking antiformants during select speech regions of interest. Potential improvements here include the use of formal statistics tests for detecting the presence of zeros within a frame prior to tracking them.

## V. CONCLUSIONS

This article has presented KARMA, a Kalman-based autoregressive moving average modeling approach to formant and antiformant tracking. The contributions of this

work are twofold. The first is methodological, with improvements to the Kalman-based AR approach of Deng *et al.* (2007) and extensions to enable antiformant frequency and bandwidth tracking in a KARMA framework. The second is empirical, with visual and quantitative error analysis of the KARMA algorithm demonstrating improvements over two standard speech processing tools, WaveSurfer (Sjölander and Beskow, 2005) and Praat (Boersma and Weenink, 2009).

It is expected that additional improvements will come with better understanding of precisely how formant information is captured through this class of nonlinear ARMA (or nonparametric) cepstral coefficient models. As noted, antiformant tracking remains challenging, although it has been shown here that appropriate results can be obtained for selected cases exhibiting antiresonances. The demonstrated effectiveness of this approach, coupled with its ability to capture uncertainty in the frequency and bandwidth estimates, yields a statistically principled tool appropriate for use in clinical and other applications where it is desired, for example, to quantitatively assess acoustic features such as nasality, subglottal resonances, and coarticulation.

## APPENDIX: DERIVATION OF CEPSTRAL COEFFICIENTS FROM THE ARMA SPECTRUM

Assume an ARMA process with the minimum-phase rational transfer function

$$T(z) \triangleq \frac{B(z)}{A(z)} = \frac{1 + \sum_{j=1}^q b_j z^{-j}}{1 - \sum_{i=1}^p a_i z^{-i}}, \quad (\text{A.1})$$

which in turn implies a right-sided complex cepstrum. For the moment, assume  $b_j = 0$  for  $1 \leq j \leq q$  to initially derive the all-pole LPC cepstrum whose  $\mathcal{Z}$ -transform is

denoted by  $C(z)$ :

$$C(z) \triangleq \log T(z) = \sum_{n=0}^{\infty} c_n z^{-n}, \quad (\text{A.2})$$

where

$$c_n = \frac{1}{2\pi} \oint_{z=e^{i\omega}} (\log T(z)) z^{n-1} dz$$

is the  $n$ th coefficient of the LPC cepstrum.

Using the chain rule,  $\frac{d}{dz} T(z)$  can be obtained independently from Eq. (A.1) or Eq. (A.2), yielding the relation

$$\frac{dC(z)}{dz^{-1}} = \frac{1}{T(z)} \frac{dT(z)}{dz^{-1}} = \frac{\sum_{i=1}^p i a_i z^{-i+1}}{1 - \sum_{i=1}^p a_i z^{-i}},$$

which implies that

$$\frac{\sum_{i=1}^p i a_i z^{-i+1}}{1 - \sum_{i=1}^p a_i z^{-i}} = \sum_{n=0}^{\infty} c_n \frac{d}{dz^{-1}} (z^{-n}) = \sum_{n=0}^{\infty} n c_n z^{-n+1}.$$

Rearranging the terms above, we obtain

$$\sum_{n=0}^{\infty} n c_n z^{-n+1} = \sum_{i=1}^p i a_i z^{-i+1} + \sum_{i=1}^p a_i z^{-i} \sum_{n=0}^{\infty} n c_n z^{-n+1}. \quad (\text{A.3})$$

Using Eq. (A.3), we can match the coefficients of terms on both sides with equal exponents. In the constant-coefficient case (associated to  $z^0$ ), we have  $c_1 = a_1$ . For  $1 < n \leq p$ , we obtain

$$c_n = a_n + \sum_{i=1}^{n-1} \frac{n-i}{n} a_i c_{n-i} = a_n + \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

On the other hand, if  $n > p$ , then Eq. (A.3) implies that

$$c_n = \sum_{i=n-p}^{n-1} \frac{n-i}{n} a_i c_{n-i} = \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

In summary, we have obtained the following relationship between the prediction polynomial coefficients and the complex cepstrum:

$$c_n = \begin{cases} a_1 & \text{if } n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if } 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if } p < n. \end{cases}$$

Reversing the roles of  $i$  and  $(n-i)$  yields the all-pole version in Eq. (5a).

To allow for nonzero  $b_j$  coefficients in Eq. (A.1), we obtain the ARMA cepstral coefficients  $C_n$  by separating contributions from the numerator and denominator of Eq. (A.1) as follows:

$$\begin{aligned} C_n &= \mathcal{Z}^{-1} \log T(z) \\ &= \mathcal{Z}^{-1} \log \frac{1}{A(z)} - \mathcal{Z}^{-1} \log \frac{1}{B(z)} \\ &= c_n - c'_n, \end{aligned}$$

yielding the respective pole and zero recursions of Eqs. (5).

- Arsikere, H., Lulich, S. M., and Alwan, A. (2011). "Automatic estimation of the first subglottal resonance", J. Acoust. Soc. Am. **129**, EL197–EL203.
- Atal, B. S. and Hanauer, S. L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am. **50**, 637–655.
- Atal, B. S. and Schroeder, M. R. (1978). "Linear prediction analysis of speech based on a pole-zero representation", J. Acoust. Soc. Am. **64**, 1310–1318.
- Boersma, P. and Weenink, D. (2009). "Praat: Doing phonetics by computer", version 5.1.40 retrieved from www.praat.org 13 July 2009.
- Broad, D. J. and Clermont, F. (1989). "Formant estimation by linear transformation of the LPC cepstrum", J. Acoust. Soc. Am. **86**, 2013–2017.
- Childers, D. G., Skinner, D. P., and Kemerait, R. C. (1977). "The cepstrum: A guide to processing", Proc. IEEE **65**, 1428–1443.
- Christensen, R., Strong, W., and Palmer, E. (1976). "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech", IEEE Trans. Acoust. **24**, 8–14.
- Deng, L., Acero, A., and Bazzi, I. (2006a). "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint", IEEE Trans. Audio Speech Lang. Processing **14**, 425–434.
- Deng, L., Cui, X., Pruvencok, R., Huang, J., Momen, S., Chen, Y., and Alwan, A. (2006b). "A database of vocal tract resonance trajectories for research in speech processing", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 369–372.
- Deng, L., Lee, L. J., Attias, H., and Acero, A. (2004). "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances", IEEE International Conference on Acoustics, Speech, and Signal Processing **1**, I-557–60.
- Deng, L., Lee, L. J., Attias, H., and Acero, A. (2007). "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model", IEEE Trans. Audio Speech Lang. Processing **15**, 13–23.
- Fulop, S. A. (2010). "Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction", J. Acoust. Soc. Am. **127**, 2114–2117.
- Garofolo, J. S., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, Philadelphia, PA).
- Hamilton, J. D. (1994). *Time Series Analysis* (Princeton University).
- Julier, S. and Uhlmann, J. (1997). "A new extension of the Kalman filter to nonlinear systems", Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation, and Controls **3**, 26.
- Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems", Trans. ASME J. Basic Eng. **82**, 35–45.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. **67**, 971–995.
- Kopec, G. (1986). "Formant tracking using hidden markov models and vector quantization", IEEE Trans. Acoust. **34**, 709–729.
- Ljung, L. (1999). *System Identification* (Upper Saddle River, NJ: Prentice-Hall).
- Marelli, D. and Balazs, P. (2010). "On pole-zero model estimation methods minimizing a logarithmic criterion for speech

- analysis", IEEE Trans. Audio Speech Lang. Processing **18**, 237–248.
- McCandless, S. (1974). "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. Acoust. **22**, 134–141.
- Mehta, D. D., Rudoy, D., and Wolfe, P. J. (2011). "Joint source-filter modeling using flexible basis functions", IEEE Int. Conf. Acoust. Speech Signal Processing .
- Miyanaga, Y., Miki, N., and Nagai, N. (1986). "Adaptive identification of a time-varying arma speech model", IEEE Trans. Acoust. **34**, 423–433.
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception", J. Acoust. Soc. Am. **85**, 2088–2113.
- Rigoll, G. (1986). "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended kalman-filter", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **11**, 1229–1232.
- Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am. **49**, 583–590.
- Rudoy, D., Quatieri, T., and Wolfe, P. (2011). "Time-varying autoregressions in speech: Detection theory and applications", IEEE Trans. Audio Speech Lang. Processing **19**, 977–989.
- Rudoy, D., Spendley, D. N., and Wolfe, P. J. (2007). "Conditionally linear Gaussian models for estimating vocal tract resonances", Proc. INTERSPEECH 526–529.
- Schafer, R. W. and Rabiner, L. R. (1970). "System for automatic formant analysis of voiced speech", J. Acoust. Soc. Am. **47**, 634–648.
- Sjölander, K. and Beskow, J. (2005). "WaveSurfer for Windows", version 1.8.5 retrieved 1 November 2005.
- Steiglitz, K. (1977). "On the simultaneous estimation of poles and zeros in speech analysis", IEEE Trans. Acoust. **25**, 229–234.
- Tourneret, J.-Y. and Lacaze, B. (1995). "On the statistics of estimated reflection and cepstrum coefficients of an autoregressive process", Signal Processing **43**, 253–267.
- Toyoshima, T., Miki, N., and Nagai, N. (1991). "Adaptive formant estimation with compensation for gross spectral shape", Electronics and Communications in Japan (Part III: Fundamental Electronic Science) **74**, 58–68.
- Yegnanarayana, B. (1978). "Formant extraction from linear-prediction phase spectra", J. Acoust. Soc. Am. **63**, 1638–1640.
- Zheng, Y. and Hasegawa-Johnson, M. (2004). "Formant tracking by mixture state particle filter", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 565–568.